CS664 Lecture #4: Maximum likelihood estimation, Markov chains

Some slides taken from:

- **Yuri Boykov**

http://www.csd.uwo.ca/faculty/yuri/Courses/WWW_433/index.html

Announcements

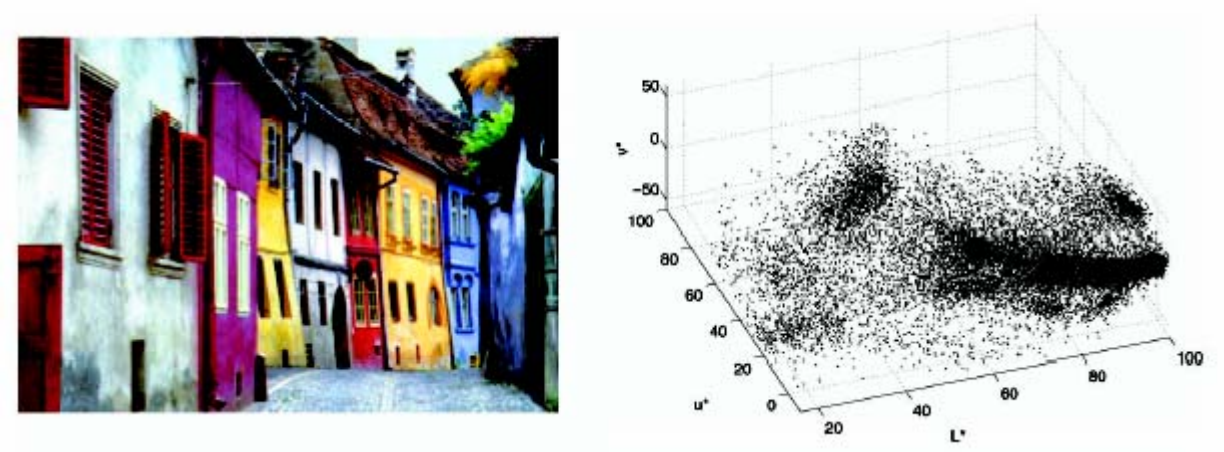
- First quiz will be on Thursday
  - Coverage through lecture #3
  - Graded via CMS, the Course Management System
  - Example questions on mean shift
    - Mean shift won't be on the quiz
- Guest lecture a week from Tuesday by Bryan Kressler
  - We'll have a few of these over the semester

Recap: Efros & Leung

- Please be sure to ask questions!
  - You will probably have to implement these
- Efros & Leung
  - Example for 1-by-1, 1-by-2 windows
    - Estimation, sampling
    - Center-weighted L_2 distance
  - Parzen estimation
    - Uniform kernel
    - Gaussian kernel

Recap: mean shift

- Mean shift in histogram space



Last lecture we saw:

- Non-parametric density estimation
  - Histograms + various fitting methods
  - Nearest neighbor
  - Parzen estimation
- Finding local modes
  - Mean shift and applications
- Maximum likelihood estimation

Probability vs. Statistics

- Probability: Mathematical models of uncertainty predict outcomes
  - This is the heart of probability
  - Models, and their consequences
    - What is the probability of a model generating some particular data as an outcome?
- Statistics: Given an outcome, analyze different models
  - Did this model generate the data?
  - From among different models (or parameters), which one generated the data?

Definition of likelihood

- Intuition: the true PDF should not make the sample (data) you saw a “fluke”
  - It’s possible that the coin is fair even though you saw 10^6 heads in a row…
- The likelihood of a hypothesis is the probability that it would have resulted in the data you saw
  - Think of the data as fixed, and try to chose among the possible PDF’s
  - Often, a parameterized family of PDF’s
    - ML parameter estimation

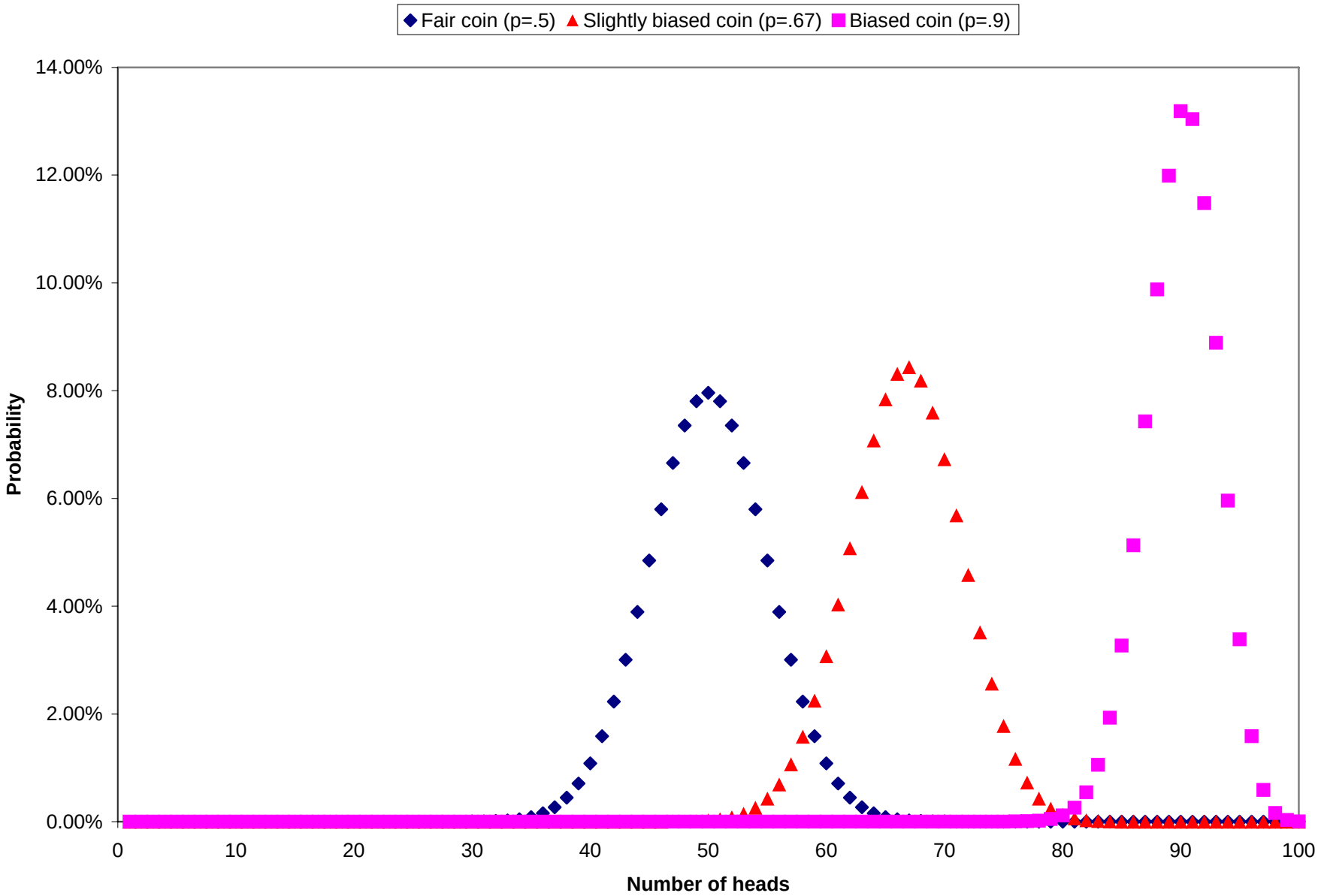
An example likelihood

- Consider a coin with probability of heads h. If we flip it n times, the probability of observing k heads is

$$\binom{n}{k} \cdot h^k (1-h)^{n-k}$$

  - Suppose we observe 51/100 heads. What is the likelihood, as a function of h?

    $\ell(h = .50) = 7.8\%$


Fair coin (p=.5)
Slightly biased coin (p=.67)
Biased coin (p=.9)
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
Probability
0
10
20
30
40
50
60
70
80
90
100
Number of heads

Coin likelihood notes

- The maximum likelihood estimate is always that the coin's bias was exactly what we saw
  - But how likely this is depends on what we saw
  - A biased coin is a better explanation for a skewed example than a less-biased coin is for a less-biased example
- Suppose we only have a few hypotheses
  - When do their likelihoods "cross"?
    - 58.5 (for $h = 0.5$ *vs.* $h = 0.67$)
    - 73.5 (for $h = 0.5$ *vs.* $h = 0.9$)

Q: What is a statistic?

- A: Anything computed from the data
  - More or less the formal definition
  - Example: sample mean
    - Percentage of heads in coin flipping
- A given model will lead to some distribution of that statistic
  - Which we just saw
- Some statistics do not allow you to select among certain models
  - Sample mean can't tell the coin has "memory"

Failure modes of ML

- Likelihood isn't the only criterion for selecting a model or parameter
  - Though it's obviously an important one
- Bizarre models may have high likelihood
  - Consider a speedometer reading 55 MPH
  - Likelihood of "true speed = 55": 10%
  - Likelihood of "speedometer stuck": 100%
- ML likes "fairy tales"
  - In practice, exclude such hypotheses
  - There must be a principled solution…

Gaussian likelihood

- With a 1-D Gaussian the probability of observing the sample $\{x_1, x_2\}$ is

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma}} \cdot \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma}}$$

  - We are assuming that both observations were drawn independently from the same distribution (i.e., same mean and variance)
    - IID (independent identically distributed)

Parametric ML example

- Suppose your sample $\{x_1, x_2, \ldots, x_n\}$ is drawn IID from a Gaussian
  - What is the ML estimate of its parameters?
  - We can maximize the log likelihood, which is

$$\log \ell(\mu) = \sum_i \log\left[\frac{1}{\sqrt{2\pi\sigma}}\right] - \frac{1}{2}\frac{(x_i - \mu)^2}{\sigma}$$

  - To maximize this we compute

$$\hat{\mu} = \arg\min_{\mu} \sum_i (x_i - \mu)^2$$

ML Parzen estimation

- What kernel maximizes the likelihood?
  - Hint: it's not very useful
- How do we make Parzen estimation actually work?
  - Can we get all possible densities as answers?
    - Do we even want to?
- Smooth kernels lead to smooth estimates
  - Choice of kernel width (often called bandwidth) is thus critical
  - Embodies an idea of what estimate we expect